Teaching Corner

Daniel J. Henderson and Christopher F. Parmeter*

Teaching Nonparametric Econometrics to Undergraduates

DOI 10.1515/jem-2015-0007
Previously published online November 6, 2015

Abstract: Given the popularity of nonparametric methods in applied econometric research, it is beneficial if students have exposure to these methods. We provide a simple, heuristic overview that can be used to discuss smoothing and nonparametric density and regression estimation suitable for an undergraduate econometrics class. We make connections to existing methods known to students (e.g. weighted least-squares through the idea of local weighting) which allows easy access to these methods. Examples are given as well as a discussion of available methods across an array of statistical software to fit the needs of educators.

Keywords: density; nonparametric; regression; teaching.

1 Introduction

In recent years high level econometric research has embraced nonparametric methods. The ability to glean information from the data while minimizing the number of parametric assumptions placed on the data generating process has clear appeal. However, the teaching of nonparametric methods still lags behind, as many popular undergraduate textbooks make little to no mention of kernel smoothing methods. Moreover, even popular graduate texts which include chapters on nonparametric estimation list these chapters as optional and course outlines rarely include this coverage or leave adequate time to detail these methods inside the classroom.

Undoubtedly, one of the main reasons for resistance to broach nonparametric methods in the econometrics classroom is the fact that, at least until recently, many popular software routines did not include automated, "black box" implementation of even basic density and regression estimators. Without this software, a heavy computing burden is placed on both the professor and the students since they need to write their own code to implement these methods, which can go well outside the scope of many undergraduate econometrics classes.[1] Further, the perceived complicated intuition of nonparametric methods also clouds the room. From our perspective, nonparametric methods are no more difficult to conceptualize than their classical parametric counterparts, once the issue of smoothing is related to students. Moreover, a simple weighted least-squares interpretation exists for kernel regression that can be easily relayed to students once discussion of weighted least-squares for the linear regression model has been covered.

Relaxing functional form specifications offers an interesting avenue to open discussion in the classroom. Consider the recent report of Collins (2012) that documents a dramatic shift in the distribution of starting

1 This is similar to the use/discussion of bootstrapping in introductory econometrics courses, see O'Hara (2014).

*Corresponding author: Christopher F. Parmeter, Department of Economics, University of Miami, Miami, FL 33146-2000, USA, E-mail: cparmeter@bus.miami.edu
Daniel J. Henderson: Department of Economics, Finance and Legal Studies, University of Alabama, Box 870224, Tuscaloosa, AL 35487-0224, USA

salaries for law graduates between 1991 and 2011. In 1991, the density was compressed with a median salary of $40,000 and three (somewhat distinct) modes, at $30,000, $50,000 and $75,000. By 2011, the density had considerably widened with a median starting salary of $60,000 and two distinct modes, at $50,000 and $160,000. An interesting classroom question is what precipitated this shift in starting salaries? More importantly, this example also serves to illustrate the pitfalls of always resorting to common parametric specifications when engaging in econometric estimation. If the data were simply fit with a normal or log-normal distribution, these modes would be completely obscured. Using nonparametric methods, the researcher has the ability to reveal these modes, which then opens the possibility for discussing why these modes appear.

The remainder of this review provides background intuition for both kernel density and regression estimation, providing details that can be quickly discerned at the undergraduate level. This leads to density and regression examples that can easily be used in the classroom. Finally, a discussion of available econometric software which allows easy access to these methods is discussed.

2 Discussing Smoothness

What is smoothness and how does it impact/influence nonparametric estimation? Given the lack of parametric assumptions, all underlying structure must be pulled directly from the data. However, without a proper notion of location, it is difficult to connect different data points as being close, or "local" to one another. In our experience, a key to discussing nonparametric methods is getting the students to understand the idea of local weighting. Once local weighting is understood, construction of a density or a regression curve is not nearly as difficult because these methods can always be linked back to similar parametric concepts.

When we think of constructing the sample mean, $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, we treat all of our observations equally. Notice that we could have written the average as $\bar{x}=\sum_{i=1}^{n}x_i/n\equiv\sum_{i=1}^{n}p_ix_i$, where $p_i=1/n$. Kernel methods, and nonparametric methods more generally, seek to change how we weight observations, based on the location of an observation to the point of interest. In the context of the unconditional mean, there is no "point" of interest and so equal weighting is appropriate. However, if we desire to know the density of individuals near a given point, then our notion of "nearness" becomes important for how we treat different observations. It is precisely this notion of nearness that smoothing attempts to capture and allows the "data to speak for themselves".

While this discussion may seem to suggest that the results of a data analysis are too dependent on the choices made by researchers and as such key insights may be driven by arbitrary choices about the degree of smoothness, we argue that there is no more arbitrariness in this process of statistical estimation than there is in parametric analysis. The choice of parametric model is typically as *ad hoc* as picking the smoothing parameter; sometimes seemingly at random in applied economics. This is because a majority of economic theories rarely yield closed form solutions about the relationship of interest. It is just as likely that key insights into an important economic relationship are driven via the choice of parametric density assumed for the data or the specification of the conditional mean in a regression study.

3 Density Estimation

To formalize our discussion on how we can construct a nonparametric estimator of the unknown density, lets first consider a crude way to approximate a density. The simplest way to nonparametrically estimate a density without assuming a specific family, such as the normal, is to use a histogram. Specifically, for a univariate random variable x with support $\mathcal{S}(x)=[a, b]$, divide $\mathcal{S}(x)$ into J equally spaced boxes. In this setup, each box will have width $h=(b-a)/J$ and we will call h our binwidth. The intervals can be described as

$$(a+(j-1)h,\ a+jh], \quad \text{for } j=1,\ 2,\ \ldots,\ J.$$

Let n_j denote the number of observations from the sample that fall in interval j. This can be summarized as

$$n_j = \sum_{i=1}^{n} \mathbf{1}\{a+(j-1)h < x_i \leq a+jh\},$$

where $\mathbf{1}\{A\}$ is the indicator function that takes the value 1 if the event A is true and zero otherwise. The proportion of observations falling into the jth interval (bin) is n_j/n. The expectation of the proportion of observations in the jth interval ($E[n_j]/n$)) is then

$$E[n_j]/n = Pr(a+(j-1)h < x \leq a+jh) = \int_{a+(j-1)h}^{a+jh} f(x)dx, \tag{1}$$

where $f(x)$ is the density of x, which is what we seek to estimate. The first equality in (1) follows by the definition of the empirical cumulative distribution function and the second is from the relationship between the cumulative distribution and the probability density function.

Now, if we assume that J is very large (h is very small), then on the interval $(a+(j-1)h, a+jh]$, we can take $f(x)$ to be approximately constant. This suggests that a crude estimator of a density function is

$$\hat{f}(x) = \frac{n_j}{nh}. \tag{2}$$

Unfortunately, our estimator is biased in finite samples, discontinuous and nondifferentiable. We would like an estimator of our density that is both consistent and smooth.

Relatively simple calculus can show that the bias is directly influenced by the binwidth (same direction) and the variance is influenced inversely by the binwidth. Further, the variance is influenced by the sample size. Thus, if we hold h fixed, while there is no improvement in the bias, the variance will decrease as the sample size increases. This conflict between controlling the bias and variance of the crude density estimator is at the heart of nonparametric estimation. There is always a delicate interplay between these two that is directly controlled via the binwidth. If we require that $nh \to \infty$ and $h \to 0$ as $n \to \infty$, then we see that both our bias and variance decrease, suggesting that our estimator converges in the mean square error sense. In essence, as we get more information about the sample (n increases), we can further partition $\mathcal{S}(x)$, thus decreasing the binwidth, which lowers bias. At the same time, we cannot lower h so much so that it offsets the gains in information that we obtain, thus $h/n \to 0$. It is common to refer to nh as the "local sample size." Roughly speaking, it is the actual amount of data that is being passed to the estimator.

When we calculate our density using this crude approximation, we take no stock of where the data are located. Thus, since our density estimate is constant in a given bin, certain points will be estimated more precisely in the bin than others. We can construct a better estimator that takes into account the location of the observations to further reduce the bias that is inherent in our current density estimator.

3.1 The Kernel Density Estimator

The histogram placed evenly spaced boxes over $\mathcal{S}(x)$, making no distinction of where the data was located. To avoid this issue, and hence improve the bias, we can instead place boxes along $\mathcal{S}(x)$ that are centered on each observation. If we label our point of interest as x, then using boxes that are centered over our sample observations, the number of observations that are within $2h$ (the centered binwidth) of our point of interest is

$$n_x = \sum_{i=1}^{n} \mathbf{1}\{x-h < x_i \leq x+h\}.$$

The corresponding probability of falling in this box (centered on x) is n_x/n. A natural estimator of our density is then

$$\hat{f}(x)=\frac{n_x}{2nh}=\frac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}\{x-h<x_i\leq x+h\},$$

where the "2" shows up because we are looking for data both to the right ($x+h$) and left ($x-h$) of x. For convenience, we can write our density estimator as

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}(1/2)\mathbf{1}\left\{\left|\frac{x_i-x}{h}\right|\leq 1\right\}. \tag{3}$$

For the discussion that follows, we are going to generalize our estimator slightly. First, we will replace $(1/2)\,\mathbf{1}\left\{\left|\frac{x_i-x}{h}\right|\leq 1\right\}$ in (3) with a "kernel" function $k(u)$ where

$$k(u)=\begin{cases}1/2 & \text{if } |u|\leq 1\\ 0 & \text{otherwise.}\end{cases}$$

Note that we have written $k(u)$ for notational convenience. In general, you should take u to be $(x_i-x)/h$, which represents how "local" the observation x_i is relative to the point of construction, x. The use of the word kernel may be stark for undergraduates; the easiest approach is simply to describe the kernel as a local weighting function. The kernel used above is known as the uniform kernel.

We see that one issue with our use of the uniform kernel is that it is discontinuous at –1 and 1 and has derivative 0 everywhere except at these two points (where it is undefined). This suggests that our density estimator will not be smooth.

We can develop a more general density estimator based off the intuition afforded from the naïve density estimator. Recall that we used the kernel function representation for the naïve density estimator, where the kernel function was equivalent to the uniform weighting function on the interval [–1, 1]. To allow for general discussion of density estimation, in what follows, we use the generic definition for a kernel density estimator

$$\hat{f}(x)=\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{x_i-x}{h}\right). \tag{4}$$

Given the generality of this definition regarding what the kernel function ($k(u)$) may look like, we call h a bandwidth as opposed to a binwidth. In general, we will select kernel functions that are much smoother than the uniform kernel. The bandwidth now corresponds to the smoothness of the estimator instead of the direct width that the kernel covers.

There are many possibilities for kernel functions, but we typically assume they possess the following properties: (1) integrate to one, (2) are symmetric and (3) have a finite variance. This is why we often use probability density functions for kernels.

Table 1 presents several of the most popular kernels employed in kernel density estimation. The biweight kernel is occasionally referred to as the quartic kernel (since it is a quartic function in u), while the Epanechnikov kernel obtained its name from Epanechnikov (1969), who showed several optimality properties for this specific kernel. The most popular kernel in econometrics is the Gaussian kernel as it has derivatives of all orders (and economists are typically interested in higher-order derivatives).

3.2 Practical Considerations

From an applied standpoint, the two most important choices facing the practitioner are the choice of kernel and the choice of smoothing parameter. A simple exercise is to present students with a single dataset and have them keep the bandwidth fixed but change the kernel. In most instances the change in the estimated density will be minimal. This invariance of the density estimator serves to demonstrate that the choice of

Table 1: Commonly Used Kernels.

Kernel	$k(u)$
Uniform	$\frac{1}{2}\mathbf{1}\{\lvert u\rvert\leq 1\}$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbf{1}\{\lvert u\rvert\leq 1\}$
Biweight	$\frac{15}{16}(1-u^2)^2\mathbf{1}\{\lvert u\rvert\leq 1\}$
Triweight	$\frac{35}{32}(1-u^2)^3\mathbf{1}\{\lvert u\rvert\leq 1\}$
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-(1/2)u^2}$

kernel has little impact on what we learn. However, a similar exercise keeping the kernel fixed but changing the bandwidth can reveal quite vividly the impact of location and smoothing. In many instances moderate changes in the bandwidth can lead to large visual changes in the underlying density.

While many may be uncomfortable with an estimator that depends heavily on the choice of a parameter that is of no real empirical interest, this is no different than pre-selection of the parametric density to fit with the data at hand. What is key is that the bandwidth is selected in such a fashion that it possesses desirable statistical properties. For example, we could use a data-driven approach such as least-squares cross-validation which is known to minimize the squared distance between the true underlying density and the estimated density. Regardless of which bandwidth is used, students can readily see the importance of the bandwidth on what can be learned about the underlying shape of the density. Further, many applied studies use kernel density estimates to set the table for a more in-depth discussion of an interesting economic phenomena. For example, Quah (1993) uses the emerging bimodality in the distribution of worldwide output to refute the widely accepted notion of economic convergence.

To illustrate the importance of the bandwidth on the statistical properties of the kernel density estimator, consider Figure 1. Here we present the mean squared error (MSE) of the kernel density estimator for various

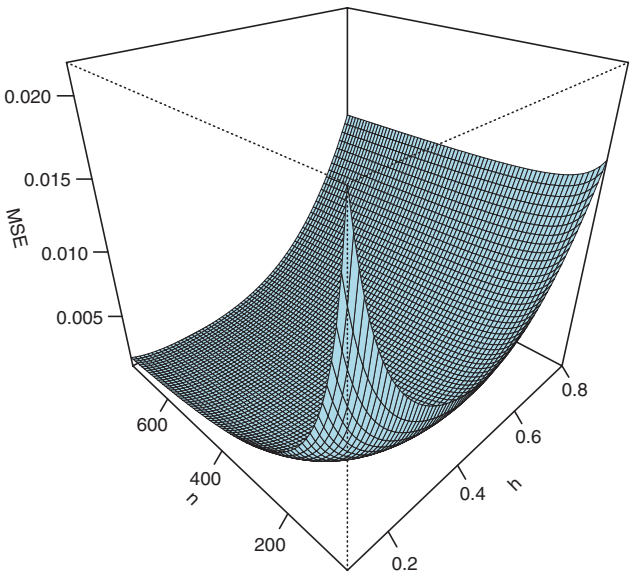

Figure 1: Mean Squared Error of the Kernel Density Estimator for Various Bandwidths and Sample Sizes when the Data are Normally Distributed.

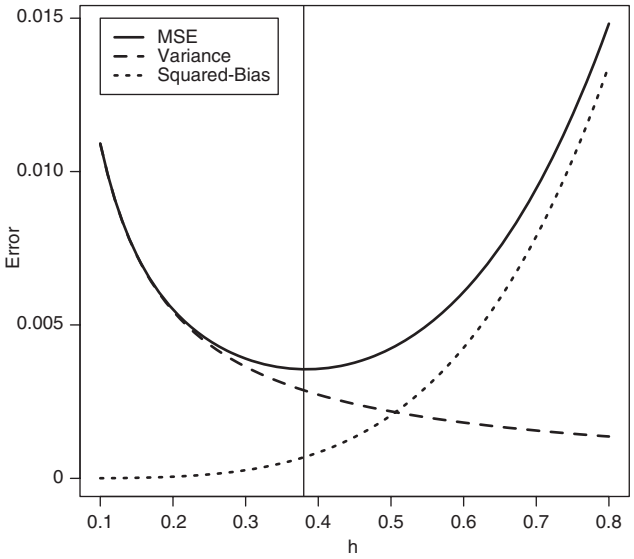

Figure 2: Mean Squared Error, Squared Bias, and Variance of the Kernel Density Estimator for Various Bandwidths with Fixed Sample Size when the Data are Normally Distributed.

combinations of sample size (n) and bandwidth (h) for data from a normal density. This figure makes it clear that for a fixed bandwidth, a larger sample size results in a decrease in MSE (working through a reduction in the variance only), while for a fixed sample size, decreasing the bandwidth can either increase or decrease MSE, given the opposite effects through the bias and variance.

This is instructive for students as it suggests that the bandwidth is not just some parameter to set, but requires careful consideration on its choice. A potential link that can be drawn is with the Gauss-Markov theorem, whereby the OLS estimator is favored in the class of linear and unbiased estimators given that it has the smallest variance. Here, because we have a bias, we must trade-off one for the other. We can also illustrate this in a similar manner by fixing the sample size, and focusing attention on a single point of the density. Figure 2 presents the MSE, squared bias and variance of the kernel density estimator for different values of the bandwidth with a fixed sample size. Here all calculations are for the point $f(0.25)$. It is clear that as h increases, the variance decreases, but the squared bias increases. Where to fix h is determined by the lowest value of MSE. Students with a good statistical background will immediately understand this example from discussions on comparing biased estimators. With two biased estimators, we cannot solely compare variances, so we look at MSE; here we are doing exactly that.

Given the importance of the bandwidth, it is clear that its choice is important. Most statistical software that constructs kernel density estimators select the bandwidth in some way that is optimal. While there are many definitions of optimality, in general, most criterion circle back to some notion of balancing between bias and variance. Instructors should be sure to emphasize to students how the software they are using selects the bandwidth. For example, in `Stata` the base `kdensity` operation uses a rule-of-thumb, which has almost no connection to optimality unless the data are actually normally distributed. In our own work, we primarily use least-squares cross-validation.

3.3 Construction of a Kernel Density

Suppose we wished to teach students how to construct a kernel density estimate. We could easily resort to any of the available software applications discussed in Section 5. However, building a kernel density estimator by hand is (relatively easy and) instructive and allows students to "get their hands dirty."

Here is a simple example in the `R` syntax. First, we construct a kernel function, in this case, the Epanechnikov, for a generic point (or vector of points) `x.i` and an evaluation point `x`:

```
kernel <- function(x.i,x,h) {
   u <- (x.i-x)/h
   return(0.75*(1-u^2)*ifelse(abs(u)<=1,1,0))
}
```

Second we need to evaluate this kernel, for a given bandwidth, for each point of interest for the unknown density, in this case $x_1, \ldots, x_n$. Given that this operation is carried out over each data point, we will embed the process in a loop:

```
fhat <- numeric()

for(j in 1:n) {
   fhat[j] <- mean(kernel(x.i,x[j],h))/h
}
```

The object `fhat` is storage for our estimates of the unknown density. The loop is required so that we evaluate the density for each observation. Lastly, if the data $x_1, \ldots, x_n$ were ordered from smallest to largest, then we could plot the x (sorted) vector against `fhat` to see the estimated density.

While we have presented `R` code, any statistical language that allows you to construct a loop and create their own function could easily be used. For example, in Excel, density estimation can easily be carried out using nothing more then row and column operations coupled with the `AVERAGE` function.

3.4 An Empirical Example – Lean Body Mass

We use the lean body mass index from the Australian Institute of Sport dataset taken from Cook and Weisberg (1994). This variable's distribution was also analyzed in Jones, Marron, and Sheather (1996). The main focus of their study was on the impact of bandwidth selection and what it revealed about the underlying structure of lean body mass. It was determined that the distribution of lean body mass had two distinct modes, which was intuitive given that it appeared there existed two distinct sub-populations in the data, one for males and one for females (see their Figure 2, page 405).

Here we show how to easily demonstrate the impact of kernel choice and bandwidth choice on what is learned from the kernel density estimator with this dataset. Figure 3 uses a bandwidth of 2.9, but switches between an Epanechnikov, Gaussian and Biweight kernel. All three density estimates are minimally different

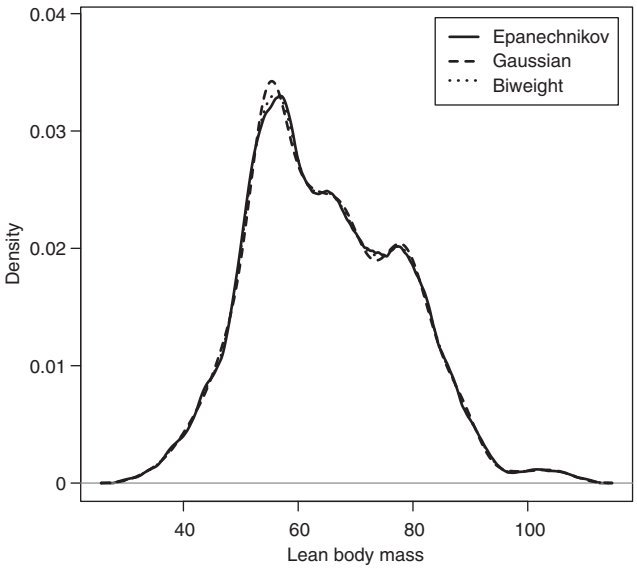

Figure 3: Sensitivity of Kernel Density Estimator to Choice of Kernel, Bandwidth Fixed at 2.9.

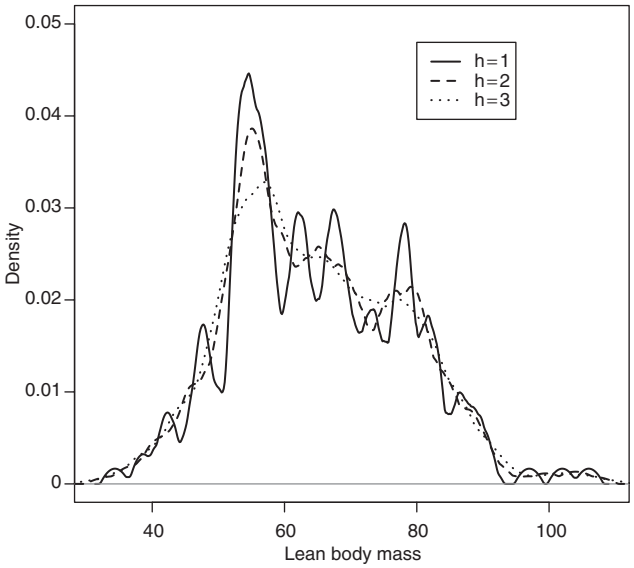

Figure 4: Sensitivity of the Kernel Density Estimator to Bandwidth, an Epanechnikov Kernel is Used in Each Case.

and evince two distinct modes. Figure 4 uses only the Epanechnikov kernel but changes the bandwidth from 1 to 2 to 3. Here what we see depends radically on the bandwidth used. The estimated density is quite irregular with many local modes using the smallest bandwidth, whereas the largest bandwidth suggest a heavily right skewed density with a single mode.

The benefit of using an example of this nature is that it can be used to discuss the importance of kernel choice (not that important) and bandwidth choice (overly important). Secondly, the reasons for the two modes in the distribution of lean body mass offers an interesting point of discussion in a classroom environment. What is it about lean body mass that might lead to such an irregular shaped density? Or, is the density actually unimodal and the features that are observed due to the selection of the bandwidth?

4 Nonparametric Regression

Regression is the backbone of applied econometric research. Although regression is widespread, the vast majority of economic research assumes that regressors enter the conditional mean linearly and that each regressor is separable without any theoretical justification. Consider the ubiquitous linear regression setup

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{5}$$

introduced at the start of an applied econometrics course. Little in the way of motivation from economic theory or practice justifies this specific relationship. Clearly, the linear relationship makes interpretation easy and estimation is straightforward. But what if this setup is incorrect? Here we discuss how to estimate regression functions where we are unsure of the underlying functional form.

The nonparametric regression estimators that we will describe here will construct an estimate of the unknown function in much the same way that we constructed the unknown density, by using a local sample for each point. Whereas parametric estimators are considered global estimators (use all data points), nonparametric kernel regression estimators are local estimators, using a local sample of nearby data points to fit a specific parametric model (typically a constant or a line) and then "smooth" each of these local fits to construct the global function estimator. This allows you to focus on the local peculiarities inherent in your data set while estimating the unknown function without judicious choice of parametric functional form.

To lay the general foundation, we assume that we have a response variable y and a single covariate which is used to predict the response, x. y and x are related through the model

$$y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where, for simplicity, we will assume that the sample realizations are distributed i.i.d. for our discussion. In general parametric analysis, a functional form for $m(\cdot)$ is assumed known up to a finite-dimensional number of unknown parameters. If we assume that $m(x_i) = \alpha + \beta x_i$, then our model can be estimated via ordinary least-squares (OLS). However, if our parametric specification of the conditional mean deviates from $\alpha + \beta x_i$, then the OLS estimator will likely be biased and inconsistent. Regardless of the exact form of $m(x)$, we can interpret it as the the conditional mean of y, $E(y|x)$. Doing so will allow us to develop a simple, nonparametric estimator for this quantity using insights from kernel density estimation.

4.1 The Local-Constant Least-Squares Estimator

Here we discuss two distinct approaches to derive a nonparametric kernel estimator of the conditional mean in the classroom. Both approaches yield the same estimator. Either approach can be used depending upon how the instructor has taught derivation of the parametric OLS estimator.

4.1.1 An Indicator Approach

If we think of constructing a naïve estimator for the conditional mean of y ($m(x)$), we could average over observations that are close to x. That is, our naïve estimator is

$$\hat{m}(x) = \frac{1}{2n_x} \sum_{i \in \mathcal{S}(x)} y_i,$$

where $\mathcal{S}(x)$ is the set of observations that are close to x and n_x is the cardinality of $\mathcal{S}(x)$. The elements of $\mathcal{S}(x)$ can be represented with a uniform kernel and the cardinality as $n_x = \sum_{i=1}^{n} \mathbf{1}\{|x_i - x| \leq h\}$. Thus, our naïve conditional mean estimator is

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i (1/2) \mathbf{1}\left\{\left|\frac{x_i - x}{h}\right| \leq 1\right\}}{\sum_{i=1}^{n} (1/2) \mathbf{1}\left\{\left|\frac{x_i - x}{h}\right| \leq 1\right\}}.$$

We can arrive at our kernel estimator by replacing the uniform kernel with the general kernel form, $k\left(\frac{x_i - x}{h}\right)$ as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i k\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)}. \tag{6}$$

4.1.2 Kernel Regression on a Constant

An alternative derivation for the nonparametric conditional mean estimator will shed light into its common name, the local-constant least-squares estimator (LCLS). We can think of estimating the unknown function ($m(x)$) as that which minimizes the weighted squared distance between the function itself and y. This is akin to weighted least-squares, except that the weights will vary by x instead of being fixed across all values of x. Think of how we construct the OLS parameter estimator. We solve

$$\min_{\alpha,\beta}\sum_{i=1}^{n}(y_i-\alpha-\beta x_i)^2,$$

by setting the first-order conditions equal to zero and obtaining the slope and intercept estimators.

If we instead replace $\alpha+\beta x_i$ with a and introduce kernel weights, we have

$$\min_{a}\sum_{i=1}^{n}[y_i-a]^2 k\left(\frac{x_i-x}{h}\right),$$

which has first-order condition

$$-2\sum_{i=1}^{n}(y_i-a)k\left(\frac{x_i-x}{h}\right)=0.$$

Solving this yields

$$a=\hat{m}(x)=\frac{\sum_{i=1}^{n}y_i k\left(\frac{x_i-x}{h}\right)}{\sum_{i=1}^{n}k\left(\frac{x_i-x}{h}\right)}$$

which is identical to (6). We essentially regress y locally, on a constant, to determine our function at a point, hence local-constant least-squares.

4.2 Connecting to Weighted Least-Squares

An elegant feature of the local constant estimator is that it can be interpreted as nothing more than a generalized least-squares estimator. In this case, the estimator is simply the regression of y on a constant, with weights that change depending upon the point of interest. This offers professors an interesting pedagogical avenue whereby students can code up a kernel regression estimator using "black box" software that performs weighted regression.

Consider how this might be done in `R`. First we construct a Gaussian kernel function for a generic point (or vector of points) `x.i` and evaluation point `x`:

```
kernel <- function(x.i, x, h) {
   u <- (x.i-x)/h
   return(dnorm(u)/h)
}
```

where `dnorm(u)` returns the value of the standard normal density evaluated at `u`.

Second we need to estimate the (nonparametric) model for each point of interest, in this case $x_1, \ldots, x_n$. We do so using the least-squares command, invoking the weights option. This is done over each data point so we embed the process in a loop:

```
mhat <- numeric()
for(j in 1:n) {
   model <- lm(y~1,weights=kernel(x.i, x[j], h))
   mhat[j] <- coef(model)
}
```

The object `mhat` is storage for our fitted values of the unknown conditional mean. Note that instead of running a single regression, as is the case with parametric regression, we run n regressions, one for each observation (i.e. evaluation point). The only thing that changes is the point which we weight over. Again,

while we have presented R code, any statistical language that allows weights to be fed into a linear regression solver can be deployed.

This approach could also be performed in Excel using the LINEST regression solver. In this case weights cannot be fed directly into the regression software. However, it is possible to construct weighted versions of y and the intercept, i.e. divide y and 1 each by $\sqrt{w_i}$, where w_i is the kernel weight for the ith observation. For each observation, repeatedly call LINEST, setting the const argument to FALSE.

4.3 Multiple Covariates

Our previous discussion was for the case of a single covariate, which keeps the notation simple, but is far from how we commonly use regression models in both the classroom and in practice. More broadly, we have access to a range of covariates. It is at this stage where the limitations of nonparametric kernel regression methods come to light; presentation issues of estimates in high dimensional settings and the curse of dimensionality. Make no mistake, kernel regression works in an almost identical fashion whether there is a single covariate or ten, one simply needs to adjust how the local weights are constructed. The most common approach would be to use what is called a product kernel, which is nothing more than successive multiplication of the individual covariate weights. Once these weights have been calculated, kernel regression would proceed in exactly the same fashion as detailed above. This makes the idea of application of the estimator simple for students. What can be more challenging is to discuss the best way to interpret one's estimates; see Li and Racine (2007) and Henderson and Parmeter (2015) for more advanced discussion in this realm.

4.4 Age-Earnings Profile

The seminal work of Jacob Mincer on human capital suggested that the logarithm of a worker's earnings is concave in their age (potential work experience). Concavity is consistent with the investment behavior implied by the optimal distribution of human capital investment over a worker's life cycle. A voluminous literature within labor economics has generally specified age-earnings profiles as quadratic (Heckman and Polachek 1974), consistent with concavity. Murphy and Welch (1990) challenged the conventional empirical strategy of specifying a quadratic in age for an age-earnings profile. Their work suggests that a quadratic specification in age understates early career earnings growth by 30–50% and overstates mid-career earnings growth by 20–50%. An analysis of residual plots from their estimated quadratic relationships (as well as several statistical tests) reveal patterns suggesting substantial differences from this specification. They advocate on behalf of a quartic age-earnings profile and find that this specification yields a substantial improvement in fit relative to the common quadratic relationship.

Given that the human capital theory of Mincer does not suggest a particular empirical relationship, Pagan and Ullah (1999, Section 3.14.2) considered the use of nonparametric regression techniques to shed light on the appropriate link between income and ages [see Basu and Ullah (1992) for a similar study]. They provided an example using the 1971 Canadian Census Public Use Tapes consisting of 205 individuals who had 13 years of education. Fitting a local-constant kernel regression function (see their Figure 3.4) they found a visually substantial difference between the common quadratic specification and their nonparametric estimates. A "dip" in the age-earnings profile around age 40 suggested that the relationship was neither quadratic nor concave. Pagan and Ullah (1999) argue that this "dip" may occur because of generational effects present in the cross-section, specifically, pooling workers who have differing earnings trajectories.

Here we use the Pagan and Ullah (1999) data and compare a linear, quadratic, and quartic parametric specification against the local-constant estimator just described. We use a bandwidth of 2.58 with the

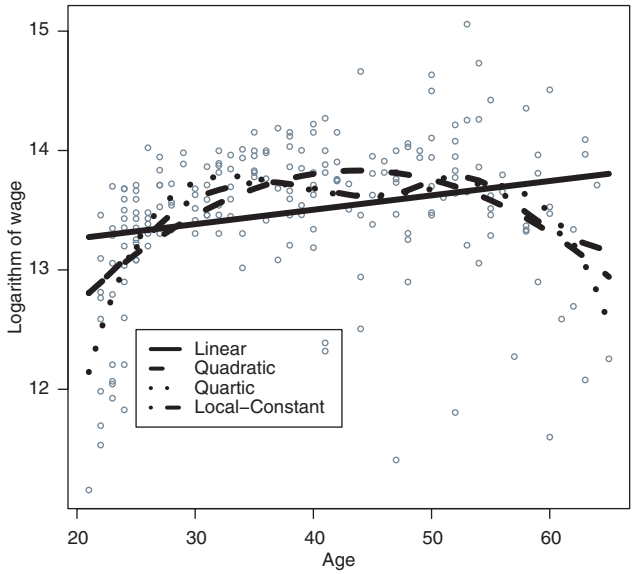

Figure 5: Age-Earning Profile for 1971 Canadian Census Public Use Tapes.

Gaussian kernel. Figure 5 plots the four different curves. It is clear that the linear relationship does not adequately capture Mincer's concave formulation. While the quadratic curve does, an interesting feature appears with both the quartic parametric specification as well as the nonparametric estimator, a dip in the relationship between earnings and age around 45. Is this a result of over-fitting or some feature of the data that simple linear and quadratic models cannot uncover?

The dip in the relationship allows for open discussion in the classroom upon demonstration. Is this dip fictional or is there some economic reason for its presence. The most common argument is that this is a cross-section and hence not a true life cycle. Regardless, the appearance of the dip allows the instructor to discuss how some parametric models can mimic the shape of nonparametric estimators and that the goal of nonparametric methods is to shed light on the potential parametric specification of a relationship and assert its robustness, not to rule out all parametric models in hand.

5 Available Software

As we discussed prior, to construct either a kernel density or kernel regression estimator, it is paramount that the software have the ability to allow users to change the bandwidth (or have automatic/data driven bandwidth selection) so that it is easily discerned the impact of the smoothing parameter on the resulting estimates. It is also useful if the software allows for kernel choice.

5.1 Stata

Within Stata, perhaps the most common undergraduate statistics interface currently used, univariate density estimation can be engaged through the `kdens` package, which leverages Stata's basic `kdensity` option. A key feature of `kdens` is that a range of kernel function and bandwidth selection mechanisms can be selected by the user. Through the `moremata` package, nine different kernel functions are available, including the Epanechnikov and Gaussian kernels.

For kernel regression, the `lpoly` function is the main conduit for nonparametric analysis. `lpoly` uses the Epanechnikov kernel as a benchmark but does allow the user to change to the Gaussian and biweight kernels (a few other options are also available). Further, a user defined bandwidth can be used. One downside to `lpoly` is that only a single covariate can be used.

5.2 SAS

The `KDE` procedure performs either univariate or bivariate kernel density estimation. Unfortunately, `PROC KDE` only uses a Gaussian kernel to compute the resulting density estimate. Data driven bandwidth selection is not currently possible directly through `KDE`, but the `BWM` option does allow the user to estimate densities using a variety of smoothing parameters.

A local-constant regression estimator does not directly exist in SAS. However, the `LOESS` procedure does allow for smoothing fitting of a curve. The `LOESS` procedure is based on the smoothing method of Cleveland and Grosse (1991), which is somewhat different than the methods described here. However, the commonly known `PROC REG` does allow weighting through the `WEIGHT` option. Here the user could construct kernel weights and then embed `PROC REG` in a loop as discussed earlier for the `R` syntax.

5.3 EViews

Kernel density estimates are straightforward to construct in EViews. From the drop down menu that appears upon selecting `View/Graph...`, simply select `Kernel Density` in the `Distribution` drop down menu. This will open up a window with many options that can be selected with the click of a mouse, including the option to use seven different kernel weighting functions. The default bandwidth is $1.059*\hat{\sigma}_x n^{-0.2}$, which is known as the Silverman rule-of-thumb bandwidth, but there is an option to provide a user specified bandwidth as well. An additional feature for students and researchers who wish to discern the impact of the bandwidth on their estimates is the `Bracket Bandwidth` option which plots three different kernel density estimates using bandwidths $0.5h$, h and $1.5h$. One unfortunate missing feature is the ability to conduct data-driven or optimal bandwidth selection within EViews.

Kernel regression in EViews works almost identically to the kernel density menu. We can perform univariate kernel regression using one of the seven kernel function options. As with density estimation, a feature that is unfortunately missing is data-driven bandwidth selection for the nonparametric regression estimator. The default bandwidth is $0.15(\max(x)-\min(x))$ and both the user specified and bracket bandwidth options are available. To access the local regression menu in EViews, simply click on `Options` and work with the `Scatterplot Customize` dialog box.

5.4 R

Out of all of the statistical software discussed in this section, `R` has perhaps the most comprehensive collection of nonparametric methods available to the user. Notably, the `np` (Hayfield and Racine 2008) and `KernSmooth` (Wand 2013) packages offer an array of cutting edge methods that can quickly implement density and regression estimation. The `np` package is the most comprehensive package of commands to engage in nonparametric analysis. For the purposes of an undergraduate lecture on either kernel density estimation or kernel regression, `np` has facilities for both data driven bandwidth selection, through `npudensbw` and `npregbw`, as well as direct estimation, through `npudens` and `npreg`. The kernel can be picked as either the Epanechnikov or Gaussian, using the option `ckertype` [see Harrison (2008) for a detailed review of all of the options available in this package].

5.5 Excel

Excel does not offer black box estimation of a kernel density or nonparametric regression function. However, there is an add-in for Excel, RExcel (http://rcom.univie.ac.at/download.html) which integrates `R`'s entire set of statistical and graphical methods into Excel.

However, a kernel density estimator is easily demonstrated for students without resorting to RExcel. It is a relatively easy exercise to have students use column and row operations on their data to construct kernel weights, which can then be copy-pasted to develop kernel weights for a range of points. The plotting facilities in Excel can then be used to plot the resulting density. Further, the `frequency` command allows histograms to be constructed with data in Excel as well.

5.6 MATLAB

In MATLAB, the Kernel Smoothing Toolbox (Horová, Kolácek, and Zelinka 2012) is available to conduct density and regression estimation quickly and efficiently. The `K_def()` command allows users to define the kernel they wish to conduct smoothing, with five different kernels (including Epanechnikov and Gaussian). To estimate a univariate kernel density, the `ksdens()` commands is all that is required.

Kernel regression can also be easily undertaken in the Kernel Smoothing Toolbox using the `ksregress()` command. The local-constant estimator can be implemented in the toolbox and data-driven bandwidth selection is also an option.

5.7 Other Software Outlets

Outside of commercial software, several other options avail themselves to engaging in nonparametric density and regression estimation. EasyReg (Bierens 2014) is freely available for download and can perform kernel density estimation and kernel regression with up to two covariates. Unfortunately, no manual for EasyReg is available, however, the software offers a variety of guides to help users become acquainted with the interface. Sephton (1998) also offers operational insights for EasyReg for interested instructors/students and reviews the software favorably. Another option is the `KDE` software of Udina (1995) which performs both kernel density and kernel regression. `KDE` is an object-oriented approach built over XLISP-STAT, which is a statistically oriented dialect of the Lisp language. The software is freely available, which again makes it a competitive alternative relative to commercial software and is thus easy to obtain distribution for a classroom environment.

6 Conclusion

This article gave an undergraduate description of nonparametric methods for density and regression estimation as well as reviewed the ability to conduct nonparametric estimation in available statistical software packages. We believe that these materials can be taught after covering weighted least-squares. The introduction of these methods should be beneficial to students continuing on to graduate study as well as those who are interested in the private sector after graduation.

Acknowledgments: We would like to thank, but not implicate, the editor, Jason Abrevaya, Darren Grant, Teresa Harrison and Michael O'Hara for valuable comments that improved the structure of the paper. We would also like to thank participants in our session at the Southern Economic Association's annual conference in Atlanta, Georgia (November, 2014) and the Eastern Economic Association's annual conference in New York City, New York (February, 2015).

References

Basu, R., and A. Ullah. 1992. "Chinese Earnings-Age Profile: A Nonparametric Analysis." *Journal of International Trade & Economic Development* 1 (2): 151–165.

Bierens, H. J. 2014. *EasyReg International*. Pennsylvania State University. http://econ.la.psu.edu/hbierens/EASYREG.HTM.

Cleveland, W. S., and E. Grosse. 1991. "Computational Methods for Local Regression." *Statistics and Computing* 1: 47–62.

Collins, J. N. 2012. "Salaries for New Lawyers: An Update On Where We Are and How We Got Here," *National Association of Law Placement Bulletin*: 1–4.

Cook, R. D., and S. Weisberg. 1994. *An Introduction to Regression Graphics*. New York: John Wiley.

Epanechnikov, V. A. 1969. "Non-Parametric Estimation of a Multivariate Probability Density." *Theory of Probability and Its Applications* 14 (1): 153–158.

Harrison, T. D. 2008. "Review of np software for R." *Journal of Applied Econometrics* 23 (6): 861–865.

Hayfield, T., and J. S. Racine. 2008. "Nonparametric Econometrics: The Np Package." *Journal of Statistical Software* 27 (5): 1–32. http://www.jstatsoft.org/v27/i05/.

Heckman, J., and S. Polachek. 1974. "Empirical Evidence of Functional Form of the Earnings-Schooling Relationship." *Journal of the American Statistical Association* 69 (346): 350–354.

Henderson, D. J., and C. F. Parmeter. 2015. *Applied Nonparametric Econometrics*. New York: Cambridge University Press.

Horová, I., J. Kolácek, and J. Zelinka. 2012. *Kernel Smoothing in MATLAB: Theory and Practice of Kernel Smoothing*. World Scientific Publishing Co. Pte. Ltd.

Jones, M. C., J. S. Marron, and S. J. Sheather. 1996. "A Brief Survey of Bandwidth Selection for Density Estimation." *Journal of the American Statistical Association* 91: 401–407.

Li, Q., and J. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. Princeton: Princeton University Press.

Murphy, K. M., and F. Welch. 1990. "Empirical Age-Earnings Profiles." *Journal of Labor Economics* 8 (2): 202–229.

O'Hara, M. E. 2014. "Pulling Econometrics Students Up by Their Bootstraps." *Journal of Economic Education* 45 (2): 121–130.

Pagan, A., and A. Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Quah, D. 1993. "Galton's Fallacy and Tests of the Convergence Hypothesis." *The Scandinavian Journal of Economics* 95: 427–443.

Sephton, P. S. 1998. "Easyreg: Version 1.12." *Journal of Applied Econometrics* 12 (2): 203–207.

Udina, F. 1995. "Interactive Graphics for Kernel Density Estimation." Unpublished Working Paper.

Wand, M. 2013. "KernSmooth: Functions for Kernel Smoothing for Wand & Jones (1995)." R package version 2.23-10. http://CRAN.R-project.org/package=KernSmooth.

Supplemental Material: The online version of this article (DOI: 10.1515/jem-2015-0007) offers supplementary material, available to authorized users.